==========

КОМПЬЮТЕРЛІК ҒЫЛЫМДАР, АСПАП ЖАСАУ ЖӘНЕ АВТОМАТТАНДЫРУ
КОМПЬЮТЕРНЫЕ НАУКИ, ПРИБОРОСТРОЕНИЯ И АВТОМАТИЗАЦИЯ
COMPUTER SCIENCE, INSTRUMENTATION AND AUTOMATION

==========

IRSTI 73.37.17
https://doi.org 10.53364/24138614_2024_33_2_7

[1] R. Anayatova[*]., [1] K. Koshekov
[1]Civil Aviation Academy, Almaty, Kazakhstan

[*]*E-mail: r.anayatova@agakaz.kz*

EVALUATION OF THE EFFECTIVENESS OF THE DEVELOPED METHOD FOR CLASSIFYING EMOTIONAL STATES THROUGH SPEECH SIGNALS

Abstract. *This study presents an innovative method for classifying emotional states through speech signals, leveraging advanced signal processing and machine learning techniques. The proposed method incorporates a multi-step approach, including feature extraction, selection, and classification. Initially, key acoustic features such as pitch, intensity, formants, and Mel-frequency cepstral coefficients (MFCCs) are extracted from the speech signals. Subsequently, feature selection techniques are applied to identify the most relevant features for distinguishing different emotional states. The classification is performed using a combination of supervised learning algorithms, including support vector machines (SVM), random forests, and neural networks.*

To evaluate the effectiveness of the developed method, a comprehensive dataset comprising various emotional speech recordings was utilized. The dataset included diverse emotional states such as happiness, sadness, anger, fear, and neutrality. The performance of the classification models was assessed using standard metrics such as accuracy, precision, recall.

Experimental results demonstrated that the proposed method achieved a high accuracy rate, outperforming existing state-of-the-art techniques. The neural network model, in particular, showed superior performance in capturing the nuances of emotional expressions in speech. Additionally, the feature selection process significantly enhanced the model's efficiency by reducing computational complexity while maintaining high classification accuracy.

In conclusion, the developed method provides a robust and effective solution for classifying emotional states from speech signals, with potential applications in fields such as human-computer interaction, mental health monitoring, and affective

computing. Future work will focus on further refining the model by incorporating more diverse datasets and exploring real-time implementation possibilities.

Keywords: *emotional state classification, speech signal, feature extraction, machine learning, neural networks, human-computer interaction, affective computing, AI.*

Introduction. The ability to accurately classify emotional states from speech signals holds significant promise for a variety of applications, including human-computer interaction, mental health monitoring, and affective computing. Emotions play a crucial role in human communication, influencing both verbal and non-verbal behaviors. Understanding and interpreting these emotional cues can enhance the responsiveness and adaptability of systems interacting with humans.

Speech, as a primary mode of communication, carries rich emotional information through various acoustic features such as pitch, intensity, and rhythm. These features can be systematically analyzed to identify underlying emotional states. Traditional methods for emotion recognition have relied heavily on manual feature extraction and simplistic classification techniques, often resulting in limited accuracy and generalizability [1].

This research aims to develop and evaluate a comprehensive method for classifying emotional states from speech signals using advanced signal processing and machine learning techniques. By integrating robust feature extraction methods with state-of-the-art machine learning algorithms, this study seeks to improve the accuracy and efficiency of emotion classification systems.

The proposed method involves a multi-step process, beginning with the extraction of key acoustic features from speech signals. These features are then processed and selected for their relevance in distinguishing different emotional states. Various machine learning algorithms, including support vector machines (SVM), random forests, and neural networks, are employed to classify the emotions based on the selected features.

The effectiveness of the developed method is assessed using a diverse dataset of emotional speech recordings, encompassing a wide range of emotional expressions. The evaluation metrics include accuracy, precision, recall, and F1-score, providing a comprehensive assessment of the model’s performance.

This introduction sets the stage for a detailed analysis of the proposed method, highlighting its potential to significantly advance the field of emotion recognition from speech. The subsequent sections will delve into the methodology, experimental results, and implications of the findings, paving the way for future research and practical applications.

Main part. 1. Comparison of the developed classification method with other machine learning algorithms

For a comparative assessment of the proposed classification method based on the use of two DCNNs, let us consider the performance of other CAL methods that have

proven themselves well in practice. For this, the following algorithms were investigated in the work:

- fully connected neural network [2, p. 117];
- logistic regression [3, p. 234];
- random forest [4];
- gradient boosting [5].

For these models, a vector of informative features *MFCC*, *melspec*, *delta*, *chroma* was used as input data. However, the values of the feature coefficients were averaged over the number of frames in the audio sample. As a result, the vector of features of objects becomes one-dimensional and contains 218 elements in its composition (39 coefficients of *MFCC*, 128 *melspec*, 39 *delta*, 12 *chroma*).

In the process of searching for the optimal parameters of these models, the following configurations of the CAL algorithms were found.

The structure of a fully connected neural network consists of four fully connected layers: an input layer with 218 neurons in accordance with the dimension of the input vector, 2 hidden by 512 neurons in each, and an output layer with seven neurons according to the number of classes. The output layer is a softmax classifier Regularization layers are located between fully connected layers: in the first, every fourth block of input data is discarded, in the second and third, every second block. The nonlinearity ReLU (2) is used as the activation function of neurons. As an optimization method Adam algorithm is used. The categorical cross-entropy acts as an error function. The proportion of correct answers *acc* is taken as a quality metric when training a network on a training subset.

In the process of selecting the hyperparameters of algorithms for logistic regression, the degree of regularization was taken to be $L_2 = 10$. For the random forest model, the number of trees is set equal to *n_estimators* = 300. When implementing the gradient boosting algorithm, the number of trees was chosen equal to *n_estimators* = 200.

For fully connected neural network and logistic regression models, a standardized estimate is applied to the training data:

$$z_x = \frac{x - \mu_x}{\sigma_x}, \quad (1)$$

where x – the element of vector of informative features;

μ_x – the average of this element over all objects in the subsample;

σ_x – standard deviation of a given element for all objects in the subsample.

The described models were implemented and trained in Python 3.7 using the machine learning libraries Scikit-learn 0.23.1 [6] and Keras 2.3.1.

Table 3.5 shows the results of comparison of the proposed classifier model, based on DCNN, with other types of considered CAL models.

Table 1. Results of comparison of the proposed method of the classification of PES by speech signal with other types of CAL models

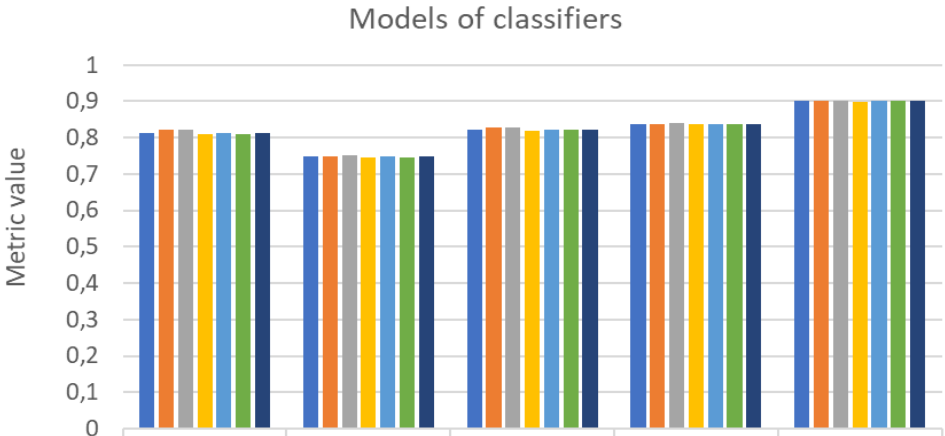
Metric type	Fully connected neural network	Logistic regression	Random forest	Gradient boosting	The proposed model of the classifier	The number of samples in the test subsample
Multi-class *acc*	0,8124	0,7494	0,8223	0,8377	0,9007	906
Average for classes *pre*	0,8223	0,7495	0,8272	0,8388	0,9017	906
Weighted average *pre*	0,8224	0,7515	0,8291	0,8402	0,9023	906
Average for classes *rec*	0,8105	0,7468	0,8199	0,8370	0,8996	906
Weighted average *rec*	0,8124	0,7494	0,8223	0,8377	0,9007	906
Average for classes *F1*	0,8106	0,7469	0,8207	0,8367	0,9001	906
Weighted average *F1*	0,8117	0,7492	0,8230	0,8378	0,9009	906

Figure 1. Presents the data from table 1 in the form of histograms.

As follows from the data obtained from the results of comparing the classifier models (table 1, figure 1), the proposed method for predicting the class of the speaker's emotional state by voice is superior to the rest of the considered CAL algorithms in all accepted types of metrics. The use of two DCNNs in the classifier model makes it possible to achieve 90% of accuracy on the test sample. The small scatter of parameters by types of metrics for the proposed classifier indicates the adequate operation of the model on seven accepted types of PES of a person. The developed classification method outperforms such effective CAL algorithms as gradient boosting and random forest in terms of performance [7].

The obtained results indicate a correctly chosen approach in the design and selection of informative features. The proposed method for detecting PES from a speech signal avoids the need to recognize said phrases in the analysis process, which greatly simplifies the classification procedure. The model uses only acoustic data for prediction. At the output of the model, the probabilities of the sample belonging to each

of the seven classes are generated. On the basis of this, it is possible to build a fuzzy logic of the operation of automatic systems for monitoring the state of a person [8].



	Fully connected neural network	Logistic regression	Random forest	Gradient boosting	The proposed model of the classifier
Multi-class acc	0,8124	0,7494	0,8223	0,8377	0,9007
Average for classes pre	0,8223	0,7495	0,8272	0,8388	0,9017
Weighted average pre	0,8224	0,7515	0,8291	0,8402	0,9023
Average for classes rec	0,8105	0,7468	0,8199	0,837	0,8996
Weighted average rec	0,8124	0,7494	0,8223	0,8377	0,9007
Average for classes F1	0,8106	0,7469	0,8207	0,8367	0,9001
Weighted average F1	0,8117	0,7492	0,823	0,8378	0,9009

Figure 1. Results of comparison of the proposed model of the classifier of the emotional state by the speech signal with other algorithms of the CAL

2. Comparison of the proposed classification model with other studies in the field

To compare the proposed method of the recognition of PES by speech signal with the research results obtained by other authors, the available sources of information in this area were analyzed. First of all, it was found that a large number of works are devoted to solving the problem of recognition of PES based on complex information about acoustic and linguistic speech data. This approach requires the existence of a special effective language model [9], which in turn significantly complicates the classification process. Moreover, it can be expected that when using aviation English with specific phraseology of radio communication, the existing language models will be ineffective. In this regard, in order to compare the research results, the classification quality metrics obtained only from the acoustic data of the speech signal were analyzed.

In addition, significant difficulties in comparing research results arise due to the use by the authors of different databases in different languages and with a different number of types of allocated PES [10].

In accordance with this, table 2 presents the results of the analysis performed comparing the quality of the classification obtained in this work, and in studies with the closest characteristics of the data used and the requirements for the results.

Table 2. The comparative analysis of the quality of the classification of the emotional state by the speech signal, obtained in this work and the similar studies

A source	Classification methods	Database	Quality metrics in %
B. Schulleretal	Gaussian Mix Model (GMM), k-Means, Support Vector Machine (SVM), Multilayer Perceptron (MLP)	B. Schuller et al.	Share of correct answers *acc* = 74,2%
C. Lee, S. Narayanan	Evaluation of the emotional weight	Private database	Classification error *err* = 1 – *acc*. *err* = 17,85% – 25,45% for men; *err* = 12,04% – 24,25% for women.
H. Goetal	Wavelet analysis, linear discriminant analysis	Private database	Accuracy *pre* = 57% – 93,3% for men; *pr*e = 68% – 93,3% for women.
M.M.H. El Ayadi et al	Gaussian mixture of vector autoregressive model (GMVAR)	F. Burkhardt at al	*acc* = 76%
B. Schuller et al	Hidden Markov Model (HMM)	Private database	*acc* = 86,8%
Javier Getal	MLP, decision trees	O. Martin et al	*acc* = 96,97%
This work	DCNN	RAVDESS; SAVEE; TESS	*acc* = 90,07%, *pre* = 90.17%
Note – Compiled from sources [8-15]			

The data in Table 2 shows that the method of proposed classification outperforms most of the known models of detection the PES from a speech signal. Moreover, in [11, 12], the share of correct answers is 96.97%, which is 6.9% higher than the results of this study. However, it should be noted that the classification by the base [13, p. 1 - 8] in the work [14, p. 20-27] was produced only for 6 types of PES without determining the neutral state. Also, in the work [14, p. 21], 264 samples of audio signals extracted from video recordings were used for the study, with one utterance for each emotion. For these reasons, it can be assumed that there is insufficient generalizing ability of those proposed in the study [14, p. 20-27] of classification algorithms.

In turn, the developed classifier based on DCNN was trained immediately on data from three different emotional corpuses (table 2), which significantly increases the generalizing ability of the final model.

Thus, based on the comparative analysis of the developed model of the speaker-independent classifier of the emotional state of a person based on his speech signal with other intelligent algorithms of CAL and proposed methods in the works of other researchers, it can be argued that the use of two DCNNs trained on the signs of mel-frequency cepstral coefficients and mel-spectrograms, is an effective solution. Moreover, the proposed classification method makes it possible to obtain a high quality of automatic detection of PES only from the acoustic data of the speech signal [12, 15].

*Conclusions.*Modern technologies of data mining make it possible to achieve high quality results in the tasks of automatically extracting useful information from various kinds of features of the objects under study. The use of deep learning technologies in the form of artificial convolutional neural networks opens up new possibilities for analyzing data of a two-dimensional structure. In particular, informative features of a speech signal have such a dimension when performing its short-term analysis to obtain mel-spectrograms, mel-frequency cepstral coefficients, and differential parameters of chalk-frequency cepstral coefficients and pitch classes.

The proposed DCNN architecture and the algorithm for its training on the selected informative features allow one to obtain high results in the classification of the emotional state of a person for seven classes of objects only on the basis of the acoustic data of the studied samples. The classifier model based on DCNN of the proposed architecture allows obtaining the best results of classification when training it on informative features in the form of mel-frequency cepstral coefficients. In this case, the result of the classification is considered as independent of the speaker, since data from three different emotional corpuses are used to train the neural network [15].

To improve the parameters of classification of PES, a method is proposed that combines the classification results from two DCNNs trained on different types of informative features: mel-spectrograms and mel-frequency cepstral coefficients. As a result, the result of classification of PES is formed in the form of the average value of the probabilities of belonging of the studied sample to each of the seven classes of PES predicted by each neural network. With this approach to solving the problem of classification of PES, it is possible to achieve a multiclass fraction of correct answers equal to 0.9007 on a deferred test subsample.

During the analysis of the results obtained, it was found that the calculated indicators of the classification quality according to the proposed method are superior to the results for other effective CAL algorithms, such as a random forest, a fully connected neural network, gradient boosting, etc. An analysis of sources based on similar studies also shows that when using only acoustic information of a speech signal to recognize seven types of PES, the proposed method surpasses the existing models in terms of quality metrics.

Р.К.Анаятова, К.Т.Кошеков

ОЦЕНКА ЭФФЕКТИВНОСТИ РАЗРАБОТАННОГО МЕТОДА КЛАССИФИКАЦИИ ЭМОЦИОНАЛЬНЫХ СОСТОЯНИЙ ПО РЕЧЕВЫМ СИГНАЛАМ

Аннотация. *В данном исследовании представлен инновационный метод классификации эмоциональных состояний по речевым сигналам, использующий передовые технологии обработки сигналов и машинного обучения. Предлагаемый метод включает в себя многоступенчатый подход, включающий извлечение, выбор и классификацию признаков. Сначала из речевых сигналов извлекаются ключевые акустические признаки, такие как высота тона, интенсивность, форманты и частотно-мелодические кепстральные коэффициенты (MFCC). Затем применяются методы отбора признаков, чтобы определить наиболее значимые признаки для различения различных эмоциональных состояний. Классификация осуществляется с помощью комбинации алгоритмов контролируемого обучения, включая машины опорных векторов (SVM), искусственного интеллекта и нейронные сети.*

Для оценки эффективности разработанного метода был использован обширный набор данных, включающий различные записи эмоциональной речи. Набор данных включал в себя различные эмоциональные состояния, такие как счастье, печаль, гнев, страх и нейтралитет. Производительность моделей классификации оценивалась с помощью стандартных показателей, таких как точность.

В заключение следует отметить, что разработанный метод представляет собой надежное и эффективное решение для классификации эмоциональных состояний по речевым сигналам, имеющее потенциальное применение в таких областях, как взаимодействие человека и компьютера, мониторинг психического здоровья и аффективные вычисления. Будущая работа будет направлена на дальнейшее совершенствование модели путем включения более разнообразных наборов данных и изучения возможностей реализации в реальном времени.

Ключевые слова: *классификация эмоционального состояния, речевой сигнал, извлечение признаков, машинное обучение, нейронные сети, человеко-компьютерное взаимодействие, эмоциональное вычисления, ИИ.*

Р.К.Анаятова, К.Т.Кошеков

СӨЙЛЕУ СИГНАЛДАРЫ БОЙЫНША ЭМОЦИОНАЛДЫ КҮЙЛЕРДІ ЖІКТЕУДІҢ ӘЗІРЛЕНГЕН ӘДІСІНІҢ ТИІМДІЛІГІН БАҒАЛАУ

Аңдатпа. *Бұл зерттеуде эмоционалды күйлерді сөйлеу сигналдары бойынша жіктеудің инновациялық әдісі ұсынылған, ол сигналдарды өңдеу мен машиналық оқытудың озық технологияларын қолданады. Ұсынылған әдіс белгілерді алуды, таңдауды және жіктеуді қамтитын көп сатылы тәсілді қамтиды. Алдымен сөйлеу сигналдарынан биіктік, қарқындылық, форманттар және жиілік-әуезді кепстральды коэффициенттер (MFCC) сияқты негізгі акустикалық белгілер алынады. Содан кейін әртүрлі эмоционалды күйлерді ажырату үшін ең маңызды белгілерді анықтау үшін белгілерді таңдау әдістері қолданылады. Жіктеу бақыланатын оқыту алгоритмдерінің, соның ішінде тірек векторлық машиналардың (SVM), жасанды интеллект және нейрондық желілердің тіркесімі арқылы жүзеге асырылады.*

Әзірленген әдістің тиімділігін бағалау үшін эмоционалды сөйлеудің әртүрлі жазбаларын қамтитын кең деректер жиынтығы қолданылды. Деректер жиынтығы бақыт, қайғы, ашу, қорқыныш және бейтараптық сияқты әртүрлі эмоционалды күйлерді қамтыды. Жіктеу модельдерінің өнімділігі дәлдік сияқты стандартты көрсеткіштермен бағаланды.

Түйін сөздер: *эмоционалды күйдің жіктелуі, сөйлеу сигналы, белгілерді шығару, машиналық оқыту, нейрондық желілер, адам-компьютерлік өзара әрекеттесу, эмоционалды есептеу, ЖИ.*

References

1. Ainakulov Z, Koshekov K, Savostin A, Anayatova R, Seidakhmetov B, Kurmankulova G (2023) Development of an advanced ai-based model for human psychoemotional state analysis. Eastern-European Journal of Enterprise Technologies, 6, 4(126), 39-49.

2. Nikolenko S., Kadurin A., Arkhangelskaya E. Deep learning. – SPb.: Peter, 2018. – 480 p.

3. Flach P. Machine learning. Science and art of building algorithms that extract knowledge from data / transl. from english A.A. Slinkina. – M.: DMK Press, 2015. – 400 p.

4. Leo B. Random Forests // Machine Learning. – 2001. – Vol. 45(1). – P. 5-32.

5. Friedman J.H. Stochastic Gradient Boosting // Computational Statistics and Data Analysis. – 1999. – Vol. 38. – P. 367-378.

6. Scikit-learn // https://scikit-learn.org/stable/. 21.09.2020.

7. Ayadi E, Kamel M, Karray F (2007) Speech emotion recognition using Gaussian mixture vector autoregressive models. In: Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference, 4, 957–960.

8. Schuller B, Rigoll G, Lang M (2004) Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture. In: Proceedings of the ICASSP 2004, 1, 577–580.

9. Schuller B (2002) Towards intuitive speech interaction by the integration of emotional aspects. In: 2002 IEEE International Conference on Systems, Man and Cybernetics, 6.

10. Lee C, Narayanan S (2005) Toward detecting emotions in spoken dialogues. IEEE Transactions on Speech and Audio Processing, 13(2), 293-303.

11. Go H, Kwak K, Lee D, Chun M (2003) Emotion recognition from the facial image and speech signal. In: Proceedings of the IEEE SICE 2003, 3, 2890–2895.

12. Koshekov K, Savostin A, Seidakhmetov B, Anayatova R, Fedorov I (2021). Aviation Profiling Method Based on Deep Learning Technology for Emotion Recognition by Speech Signal. Transport and Telecommunication, 22(4), 471-481.

13. Burkhardt F, Paeschke A, Rolfes M, Sendlmeier W, Weiss B (2005) A database of German emotional speech. In: Eurospeech, 9th European Conference on Speech Communication and Technology, 1517–1520. 14. Javier G., Sundgren D. et al. Speech emotion recognition in emotional feedback for Human-Robot Interaction // International Journal of Advanced Research in Artificial Intelligence. – 2015. – Vol. 4, №2. – P. 20-27.

15. Martin O, Kotsia I, Macq B, Pitas I (2006) The eNTERFACE'05 Audio-Visual Emotion Database. In: Data Engineering Workshops, Proceedings. 22nd International Conference.

Анаятова Разиям Курванжановна	Доктор философии (Ph.D), ассоциированный профессор кафедры «Авиационная техника и технологии» Академия Гражданской Авиации, г. Алматы, 050039, РК, e-mail: r.anayatova@agakaz.kz
Анаятова Разиям Курванжановна	Философия докторы (Ph.D), Авиациялық техника және технологиялар кафедрасының қауымдастырылған профессоры, Азаматтық Авиация Академиясы, Алматы қаласы, 050039, ҚР, e-mail: r.anayatova@agakaz.kz
Raziyam Anayatova	Ph. D, Associate professor of Aviation Engineering and Technology Department, Civil Aviation Academy, Almaty, 050039, The Republic of Kazakhstan, e-mail: r.anayatova@agakaz.kz
Кошеков Кайрат Темирбаевич	д.т.н., профессор кафедры «Авиационная техника и технологии» Академия Гражданской Авиации, г. Алматы, 050039, РК, e-mail: k.koshekov@agakaz.kz
Кошеков Кайрат Темирбаевич	т.ғ.д., Авиациялық техника және технологиялар кафедрасының профессоры, Азаматтық Авиация Академиясы, Алматы қаласы, 050039, ҚР, e-mail: k.koshekov@agakaz.kz
Kairat Koshekov	Doctor of engineering, Professor of Aviation Engineering and Technology Department, Civil Aviation Academy, Almaty, 050039, The Republic of Kazakhstan, e-mail: k.koshekov@agakaz.kz